Call for Master Thesis

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Dataless Short Text Classification for German Language

Are you interested in making a big impact with your thesis? Work with us on an innovative approach for short text classification.

Short text categorization is an important task due to the rapid growth of online available short texts in various domains such as web search snippets, short messages etc. Recently, several supervised learning approaches have been proposed for short text classification. However, most of them require a significant amount of training data and manually labeling such data can be very time-consuming and costly. Another characteristic of existing approaches is that they all suffer from issues such as data sparsity, and insufficient text length. Moreover, due to the lack of contextual information, short texts can be highly ambiguous. Thus, short text classification is much more challenging in comparison to traditional long documents. Further, if the short text to be classified is not English text, the classification task gets even more challenging, because most of the the available resources on the Web such as text classification benchmarks are in English.

In this thesis, to overcome the mentioned challenges, first we will adopt an already proposed probabilistic approach[1] to German short text classification problem. The approach does not require any labeled training data. It is able to capture the semantic relations between the entities represented in a short text and the predefined categories by embedding them into a common vector space using the recent network embedding techniques. Finally, the category of the given text can be derived based on the semantic similarity between entities present in the given text and the set of predefined categories. The similarity is computed based on the vector representation of entities and categories.
After applying the proposed approach to German short text, the final aim of the thesis would be to improve the performance of the classification task.

This thesis will be supervised by **Prof. Dr. Harald Sack, Information Service Engineering at Institute AIFB, KIT, in collaboration with FIZ Karlsruhe**.
[1] https://bit.ly/2UbNKRP

Which prerequisites should you have?

- Good programming skills in Python or Java
- Interest in Natural Language Processing
- Interest in Semantic Web technologies
- Interest in Deep Learning technologies

Contact person:
Rima Türker
rima.tuerker@kit.edu
rima.tuerker@fiz-karlsruhe.de

aifb
Institute of
Applied Informatics and
Formal Description Methods
http://www.aifb.kit.edu/

Karlsruhe Institute of Technology